MapReduce for Twitter Data Search

October 24, 2023

1 Introduction

This document outlines the MapReduce process for searching tweets containing the keyword "black friday" stored in DocumentDB on HDFS.

2 Phases

I. **Map Phase**

- *Input*: Raw tweet documents.
- *Operation*: Create inverted index entries for "black friday".
- *Output*: Key as "black friday", value as (tweet ID, metrics).

II. **Partitioning and Shuffling**

- Implement custom partitioners to ensure all records with the same keyword go to the same reducer.

III. **Reduce Phase**

- *Input*: Sorted or grouped key-value pairs from mappers.
- *Operation*: Use priority queue to maintain top 10 tweets based on a chosen metric (e.g., likes, retweets).
- *Output*: Top 10 tweets and a single count integer representing the total number of tweets with "black friday".

IV. **Optimization**

- Data Locality: Run mappers close to the data they process.
- Combiners: Use local aggregation before the shuffle phase.
- Caching: Cache frequently accessed data.